

Discovery Environments

A 'Discovery Environment' is a software system that allows Grand Challenge team participants to access the relevant data sets, integrate across them to identify connections, visualize them in ways that allow the 'big picture' to appear, manipulate the data with analytic tools, and share results by facilitating computational steering. Grand Challenge teams will work collaboratively with iPlant Collaborative staff to design and develop custom Discovery Environments tailored to help the team address a Grand Challenge question. Our model for Discovery Environments are Internet 'mashups', also known as Web 2.0 applications, which allow community members to build content in a democratic way, to make and label connections between different types of content, and to integrate a variety of different types of information in a single user interface. The wildly successful [Wikipedia project](#) is one such application. It allows users to create and interconnect knowledge using the shared metaphor of an encyclopedia. Google Maps is another well-known mashup application. It provides the community a common reference system, a detailed geographic map of the world, and encourages people to link in their own data sets indexed by GPS location. Data sets contributed by independent groups, for example average housing prices compiled by one group and mean SAT scores of students enrolled in school districts compiled by another, become dramatically more useful when linked together by a common coordinate system. Mashups reveal new patterns among data and allow one to make hypotheses of causality that would be impossible if the data sets were examined separately.

Discovery Environments are the cyber equivalent of the iPlant Collaborative physical meeting spaces. During the formative phases, they will provide a way to exchange ideas and prototypes and collaboratively create and refine the approach. During the production phase, they will provide a collaborative environment in which to exchange ideas, integrate data sets, share protocols and explore algorithmic approaches. Ultimately, they will be a way to publish the project's research findings to the world and to invite participation from the wider community.

The Integrated Solutions team will create Discovery Environment mashups for plant biology by (1) identifying long-lived data sets that will serve as shared coordinate system frameworks for integrating disparate data sets and (2) providing the community with software services that enable the layering of data sets on top of these frameworks, in a distributed, community-controlled manner. The particular data set frameworks identified by Integrated Solutions staff will depend on the Grand Challenge projects that are chosen by the community, but illustrative examples include annotated genomes, named sets of genes, their aliases and cross-species orthology relationships, phylogenetic trees, protein structures, anatomical descriptions of plant tissues and/or developmental stages, annotated collections of microscopic images, machine-readable descriptions of biochemical or regulatory pathways, and geographical descriptions of species distributions. For example, for a Grand Challenge project that requires extensive cross-species gene comparisons, we might build a Homology Registry that allows community members to assert (and dispute) phylogenetic relationships among members of gene families based on different types of evidence such as sequence conservation and synteny. For a Grand Challenge project that involves dissection of signal transduction pathways, we might provide a WIKI-like environment that allows Grand Challenge team members to assemble a comprehensive description of plant G-protein coupled receptor kinases that combines written text with embedded media that show the position of the kinases on several genomes, the evolutionary trees that relate the kinases across species, kinetic models of signaling cascades driven by the family, and a map showing the geographical distribution of allelic variants of plant kinases.

Whenever possible, Discovery Environments will be based on existing software products and will be coordinated with groups performing similar work. For example, if a Discovery Environment requires a common coordinate system based on an ontology, we will use an existing ontology such as the Plant Ontology (Ilic et al. 2007), if feasible, and coordinate the effort with the National Center for Biomedical Ontologies. Likewise, Discovery Environments based on a genome assembly and annotation will leverage interfaces developed by existing repositories such as TAIR (Garcia-Hernandez et al. 2002), Gramene (Jaiswal et al. 2006), and NCBI (Wheeler et al. 2007), rather than attempt to replace the functionality of those resources. There are numerous efforts in the bioinformatics and broader web development communities to create mashup systems, several of which would make good foundations for specific Discovery Environments, including QEDWiki (<http://services.alphaworks.ibm.com/qedwiki/>), the Taverna workflow management system (Oinn et al. 2004), AJAX GBrowse (<http://biowiki.org/view/GBrowse/WebHome>; for genome-based collaboration), and Galaxy2 (Giardine et al. 2005).